UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

AN EXAMPLE OF CORRELATION AMONG RESIDUALS IN DIRECTLY ORDERED DATA

by

Ulrich C.H. Blum* and Marc J.I. Gaudry**

* Institut für Wirtschaftspolitik und Wirtschaftsforschung, Universität Karlsruhe, P.O. Box 6980, D-7500 Karlsruhe, West Germany.

** Département de sciences économiques et Centre de recherche sur les transports, Université de Montréal, C.P. 6128, succursale A, Montréal, Canada H3C 3J7.

The research for this paper was supported by the F.C.A.R. program of Québec, the S.S.H.R.C. and the N.S.E.R.C. of Canada and the D.F.G. of Germany. Parts of Gaudry's and Blum's contributions were made during their tenure respectively as Alexander von Humboldt Research Scholar in 1985 at the University of Karlsruhe and as Alexander von Humboldt Scholar in 1986/87 at the Université de Montréal. The authors thank Anne Mouls for computing assistance.

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Universität Karlsruhe
Institut für Wirtschaftspolitik und Wirtschaftsforschung
Discussion Paper 3/86

Université de Montréal
Centre de recherche sur les transports - Publication #456
Département de sciences économiques - Cahier #8619
June 1986 - Revised December 1986

Dépôt légal - 1er trimestre 1987

ABSTRACT

We distinguish between **naturally** ordered data and data that are ordered **directly** by the researcher. After ordering observations according to a behavioral criterion, we estimate a single-order correlation parameter of the residuals simultaneously with the other parameters and the functional form of a savings equation for higher-income white-collar German households.

Key words : naturally ordered data, directly ordered data, correlation of residuals, household savings, Federal Republic of Germany.

RÉSUMÉ

Nous distinguons entre les données naturellement ordonnées et celles dont l'ordre est dirigé par le chercheur. Nous formulons un schéma de corrélation d'ordre simple entre des observations ordonnées suivant un critère économique; nous estimons ce paramètre en même temps que les autres paramètres et la forme fonctionnelle d'une équation expliquant l'épargne des ménages (de cols blancs à revenu élevé) en République Fédérale Allemande.

Mots-clés : données en ordre naturel, données en ordre dirigé, corrélation des résidus, épargne des ménages, R.F.A.

ZUSAMMENFASSUNG

Wir unterscheiden Daten in natürlicher und solche in durch den Forscher festgelegter direkter Ordnung. Nach Anordnung der Beobachtungen einer Haushaltsstichprobe über deutsche Angestellte mit gehobenem Einkommen entsprechend einem festgelegten Verhaltenskriterium zur Schätzung einer Sparfunktion untersuchen wir deren Residuen gleichzeitig mit der Bestimmung der Koeffizienten und der funktionalen Form auf Autokorrelation erster Ordnung.

Schlüsselworte : natürlich geordnete Daten, direkt geordnete Daten, Korrelation der Residuen, Haushaltsersparnis, Bundesrepublik Deutschland.

CONTENTS

1. Introduction 1

2. Residue impact criteria. 2

A. Naturally ordered data. 3

B. Directly ordered data 4

3. Example : higher-income white-collar German household savings. . . 6

4. Conclusion 11

Bibliography. 12

- - - - - - - - - - - -

Table 1 : Savings function of higher-income white-collar German households in 1978 : Elasticities, t-statistics and other results. 8

Table 2 : Savings function of higher-income white-collar German households in 1978 : Elasticities and partial derivatives. 10

1. INTRODUCTION

In recent years, increased attention has been paid to the extraction of systematic information that remains in the residual term of a regression equation. This is normal both from a numerical point of view, as the neglect of information appears inefficient, and from the statistical point of view because the useful assumption that the distribution of the residuals is spherical should be tested and not assumed a priori if many tests carried out on the results are to be credible. In this note, we draw attention to a particular new orientation of the analysis of correlation among residuals that may prove as fruitful as established orientations using naturaly ordered data. In order to distinguish the analysis of data directly ordered (DO) by the researcher from that of naturally ordered (NO) data, we consider the very simple case of a standard linear regression model to which we add a simple error process :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v \quad , \tag{1}$$

$$v = \rho R v + w \quad , \tag{2}$$

where y, X_1 and X_2 are $T\times 1$ vectors of observations, v and w are $T\times 1$ vectors of residuals, the β_k and ρ are parameters and R is the $T\times T$ **residue impact matrix** where element r_{ti}, for instance for the t^{th} row

$$v_t = \rho \sum_i r_{ti}\, v_i + w_t, \quad t=d_1+1,\ldots,n-d_2 \quad , \tag{3}$$

is defined as

$$r_{ti}\begin{cases} = 0 \quad , & t=i, \\ = \text{value established by an impact criterion}, & t\neq i. \end{cases} \tag{4}$$

For simplicity, we shall neglect observations d_1 and d_2 arbitrarily regrouped here at the "beginning" and "end" of the otherwise unordered sample.[1] In the simple case of single-order correlation considered here, only one ρ is used for all v_t elements retained but, in multiple-order correlation, this constraint can be relaxed.

Clearly, the r_{ti} elements of the residue impact matrix describe the interaction among residuals. Our task here is to comment on various approaches to the specification of residue impact criteria (RIC) required to define (4). We then proceed to give an empirical example with directly ordered data, using maximum likelihood estimation to establish the usefulness of the additional single-order correlation parameter.

2. RESIDUE IMPACT CRITERIA

We assume throughout that the sample on y and the X_k contains no missing observations. In time series, the meaning of this assumption is obvious; in spatial data, it requires that no zone of the area considered be neglected; in usual cross-sectional data, the sample need not be full but must be truly unbiased or representative of the underlying population distribution in the sense that all observed frequencies must have the same reliability or precision. Although not always necessary for the specification of certain impact criteria, this last requirement is crucial if the eventual ordering of values is to yield correct relative positions or distances among observations. Such **completeness** requirements are easily

---

[1] When the issue arises, retaining these observations complicates exposition and computations. Moreover, the "first" and "last" observations can easily create outlyers that will have excessive and unplausible effects on the results. In time series, the assumption implied by keeping these potential outlyers, namely that the first order autoregressive scheme always existed, is also too strong : one is often satisfied with the alternate view that the model makes a local approximation of a scheme that starts with the second observation.

violated in practice, notably in individual cross-sections where the degree of bias is difficult to know.

A. Naturally ordered data

Of course, time series and spatial data are "naturally" ordered but, as we shall see, the benefits that arise from this fact are limited because there is just so much that one can do by exploiting the running index of the observations.

It is convenient to associate the natural order of time series observations with their running index. Then, each row vector r_t of the residue impact matrix has the familiar appearance

$$\begin{array}{lc} & 1,\ldots,t-1,\ t,\ t+1,\ldots,T \\ r_t = & (0 \ \ldots \ 1 \ \ 0 \ \ 0 \ \ldots \ 0) \\ r_{t+1} = & (0 \ \ldots \ 0 \ \ 1 \ \ 0 \ \ldots \ 0) \end{array} \quad , \qquad (5)$$

if the residue impact criterion is

[Each residual is correlated with its previous value] . (NORIC-1)

Because NORIC-1 assumes that later observations have no influence on the present, there can only be zeroes on the right-hand-side of diagonal elements of (5). The visual and computational benefits of matching the time sequence and running indexes are significant : they are exploited by the majority of regression packages.

If spatial data are used, one would naturally wish to take into account the likely correlation between neighboring zones. As a zone can have a variable number of contiguous neighbors, a residue impact criterion such as

[Each residue is correlated with its contiguous neighbors] (NORIC-2)

might well yield rows of the impact matrix such as

$$\begin{array}{lll} & & t \qquad\qquad T \\ r_t & = (0 \ldots 0\ 1\ 0\ 1\ 0\ 0 \ldots 0\ 0\ 1) & , \\ r_{t+1} & = (0 \ldots 0\ 0\ 1\ 0\ 1\ 0 \ldots 0\ 1\ 1) & , \end{array} \tag{6}$$

if the t^{th} zone is a common neighbor of adjacent zones t-1, t+1, and T but zone t+1 has more borders than zone t (namely t, t+2, T-1, and T). Clearly, the single-order context adopted on purpose here would be a special case of more credible situations in which one might take the orientation of the zones into account (west of, south of, etc.) in order to model the influence of prevailing winds or other factors incompletely specified among the X_k regressors in (1). For greater realism, one could also make the impact a continuous function of an economic interaction variable, such as travel distance or cost, as in Bolduc (1985) who also summarizes this literature, and greatly improves the realism of the model.[2] Although the data are ordered in space, there is in general, for (6), no simple way to use the running index to order the observations in a way that fits the ordering required by a residue impact criterion such as NORIC-2.

B. Directly ordered data

As soon as the residue impact criterion requires more than searching for statistically significant leads and lags among the running index values, the modeling dimension involved in defining residue impact criteria has to be made explicit : making correlation a function of explicit variables, as just mentioned, is an example. Of course, the fact that the data are "naturally" ordered in space has in practice no more relevance

---

[2] It can be noted that, in this case, no observation is lost. If one required that only zones which are surrounded by other zones be included, then the outside zones would be "lost" as observations.

than the impact criterion will allow : if the chosen impact criterion is independent from spatial location, one can hardly say that the ordering is **natural**. In effect, the ordering becomes **directly** defined by the impact criterion used to test a specific economic hypothesis. This can be seen most easily with a cross section of individuals. Assume for expository purposes that we are interested in explaining a particular dimension of behavior, say individual (or group) savings. There are many ways in which a residue impact criterion can be specified.

[To test leadership, assume that all residuals are positively correlated with that of individual t*] . (DORIC-1)

This implies a residue impact matrix where the (non-diagonal) elements of the column t* are 1 and all others are 0. By contrast, a negative value of the correlation coefficient could be interpreted as opposition to individual t*.

More often, one will be interested in phenomena of imitation (looking up to) or distanciation (looking down on) with respect to individuals (or groups) differing in some way or other. Such ordering criteria might be defined with respect to any variable, including factors not used in the regression model. For instance,

[Rank individuals in decreasing order of net income and assume that each residual is positively correlated with that of the closest individual having 1000 units more of annual net income] . (DORIC-2)

As a rule, directly ordered data resulting from behavioral residue impact criteria will not permit easy use of the running index. The analysis of correlation is transformed into a primarily economic modeling problem and the "economic" distance between observations becomes in principle more complicated than the simple distance of time series analysis : one might, for instance, define a correlation which depends on a function of many variables!

3. EXAMPLE : HIGHER-INCOME WHITE-COLLAR GERMAN HOUSEHOLD SAVINGS

To illustrate the usefulness of directly ordering data, consider the following household savings function

$$SAV_t = f(SPI_t, PER_t, AGE_t, INT_t, TAX_t, GIN_t) + u_t \qquad (7)$$

where, neglecting subscripts (units in parentheses):

SAV = increase in net wealth (including housing but excluding durables, DM),

SPI = compulsory contributions to the public (old age) Social Security System (DM),

PER = number of persons permanently living in the household,

AGE = age of household head (years),

INT = rate of return on household portfolio (% ÷ 100)

TAX = income, capital and wealth taxes paid (DM),

GIN = gross household income including transfers (DM).

The available data, pertaining to German households in 1978, are published by the German Federal Statistical Office (1982) for various types of households, notably white-collar households. For a number of reasons explained in Blum (1986), the two parts of the stratified sample data, available by household size and age of the household's head (of the same population), were merged [3] and observations repeated to match their known share in the population. For reasons related to income tax brackets and social security contribution limits, we retained households earning a gross monthly income of 4000 DM or more.

---

[3] The missing variable (AGE or PER) in each subset of the data was replaced by the average value of the variable in the other subset. A dummy variable used to detect a resulting bias had a zero coefficient and was removed.

As we have no strong a priori views on the functional form of this function, we use the following econometric specification, outlined in Gaudry and Dagenais (1979)

$$y_t^{(\lambda_y)} = \beta_0 + \sum_k \beta_k \, x_{k_t}^{(\lambda_{xk})} + u_t \quad , \tag{8}$$

$$u_t = [\exp(\delta_0 + \delta_m z_{m_t}^{(\lambda_{zm})})]^{\frac{1}{2}} v_t \quad , \tag{9}$$

$$v_t = \rho \, v_{t-26} + w_t \quad , \tag{10}$$

where (λ_y), (λ_{xk}) and (λ_{zm}) denote Box-Cox transformations applied to each variable. In this case, **gross income** is used to obtain homoskedastic errors and the lag t-26 is chosen according to the following residue impact criterion stressing the existence of a minimum economic distance, below which households do not notice behavioral differences:

[Assume that every household "looks up" to households 7 percentiles higher in the distribution of disposable income] . (DORIC-3)

The observations were ordered according to **disposable income** and 7 percentile points (corresponding to 26 consecutive observations) was the minimum number of observations that never involved looking up to households in the same sample stratum. Because of the presence of repeated observations, the Durbin-Watson statistic, sometimes usable (King and Evans, 1985), is not helpful. We will in any case use the likelihood ratio test.

The first part of Table 1 contains elasticities of E(y), evaluated at the mean value of the explanatory variables, t-statistics conditional upon the optimal values of the Box-Cox transformations found in the second part of the table and, under the parentheses of the t-statistics, a number identifying the associated Box-Cox transformation. The second part of the

Table 1

Savings function of higher-income white-collar German households (1978): Elasticities, (t-statistics) and other results

ELASTICITIES (T-STATISTICS)	CODE NO. = DEP.VAR. =	1 WHSAV	2 WHSAV	3 WHSAV	4 WHSAV	5 WHSAV	6 WHSAV	7 WHSAV
WH = WHITE COLLAR WORKERS (HIGH)								
PAYMENTS TO SOC.PENS.INSURANCE	WHSPI	.042 (3.04) LAM 1	-.011 (-.58) LAM 1	-.013 (-.45) LAM 1	-.018 (-.65) LAM 1	.023 (.94) LAM 1	-.024 (-.80) LAM 1	.025 (.99) LAM 1
HOUSEHOLD SIZE	WHPER	.103 (2.35) LAM 1	-.166 (-3.50) LAM 1	-.165 (-2.87) LAM 1	-.151 (-2.50) LAM 1	-.106 (-3.86) LAM 1	-.149 (-2.25) LAM 1	-.100 (-3.62) LAM 1
AGE OF HOUSEHOLD'S HEAD	WHAGE	-.270 (-6.57) LAM 1	-.399 (-10.13) LAM 1	-.400 (-9.57) LAM 1	-.391 (-9.55) LAM 1	-.384 (-8.65) LAM 1	-.383 (-8.86) LAM 1	-.356 (-8.95) LAM 1
AVG. RETURN ON HOUSEHOLD PORTFOLIO	WHINT	-.649 (-3.74) LAM 1	-.054 (-.34) LAM 1	-.051 (-.45) LAM 1	-.100 (-.91) LAM 1	.260 (1.93) LAM 1	-.055 (-.50) LAM 1	.329 (2.50) LAM 1
INCOME AND CAPITAL TAXES	WHTAX	.655 (5.45) LAM 1	-.359 (-2.50) LAM 1	-.356 (-2.24) LAM 1	-.309 (-1.80) LAM 1	-.229 (-3.91) LAM 1	-.324 (-1.75) LAM 1	-.221 (-4.07) LAM 1
INCOME (BEFORE TAXES)	WHGIN	1.080 (8.37) LAM 1	2.640 (11.85) LAM 1	2.630 (9.95) LAM 1	2.560 (9.14) LAM 1	2.370 (22.94) LAM 1	2.600 (8.69) LAM 1	2.340 (23.82) LAM 1
REGRESSION CONSTANT	CONSTANT	-.732 (-2.76)	-2.000 (-6.35)	-2.030 (-5.85)	-1.680 (-4.94)	-7.230 (-17.19)	-1.720 (-4.48)	-8.690 (-18.10)
PARAMETERS (T-STATISTICS)	**CODE NO. = DEP.VAR. =**	**1 WHSAV**	**2 WHSAV**	**3 WHSAV**	**4 WHSAV**	**5 WHSAV**	**6 WHSAV**	**7 WHSAV**
AUTOCORRELATION								
ORDER 26	RHO 26						.132 (1.79)	.158 (1.98)
BOX-COX TRANSFORMATIONS								
LAMBDA(V)	LAM V	1.000 FIXED	0.000 FIXED	-.024 (-.23) (-9.63)	-.004 (-.03) (-8.93)	-.401 (-2.40) (-8.40)	-.002 (-.02) (-9.11)	-.437 (-2.70) (-8.88)
LAMBDA(X) - GROUP 1	LAM 1	1.000 FIXED	0.000 FIXED	-.024 (-.23) (-9.63)	.077 (.35) (-2.97)	-1.028 (-2.70) (-5.32)	.070 (.21) (-2.84)	-1.156 (-3.09) (-5.76)
HETEROSKEDASTICITY								
DELTA COEFFICIENTS								
INCOME (BEFORE TAXES)	WHGIN					-.035 (-7.67)		-.074 (-8.28)
BOX-COX TRANSFORMATIONS								
INCOME (BEFORE TAXES)	WHGIN					2.519 (2.26) (1.36)		2.165 (2.10) (1.13)
LOG-LIKELIHOOD		-364.138	-285.497	-285.438	-285.353	-286.103	-282.537	-282.111
PSEUDO-R2		.982	.976	.976	.976	.978	.977	.979
NUMBER OF OBSERVATIONS		341	341	341	341	341	341	341
SAMPLE -FIRST OBSERVATION -LAST OBSERVATION		27 367	27 367	27 367	27 367	27 367	27 367	27 367
NUMBER OF INDEPENDENT VARIABLES		7	7	7	7	7	7	7

table contains autocorrelation and heteroskedasticity parameters and t-statistics, as well as values of the maximized likelihood obtained with the L-1.1 algorithm (Liem et al., 1983). For our model without correlation, the values of the log likelihood for homoskedastic linear and logarithmic specifications were -364.138 and -285.497 respectively. Columns 3 and 4 respectively show that the use of one or two Box-Cox transformations — one on the endogenous and one on all exogenous variables — did not yield significant gains over the logarithmic case shown in Column 2. However, taking heteroskedasticity (according to equation 10) into account in Column 5 shows both a large gain in the log likelihood and significant differences between the two Box-Cox estimates. Column 6 shows that correcting for autocorrelation under the assumption of homoskedasticity also improves on the result of Column 4. Column 7, the most general model, shows the effet of simultaneously correcting for heteroskedasticity and autocorrelation. Compared to the results of Column 5, we note that the additional autocorrelation parameter is significant even when heteroskedasticity is taken into account. This suggests that the additional parameter, which reflects the extent to which households look up to higher income households in their savings behavior, is a useful addition to the model.

Table 2 contains elasticities and partial derivatives of savings with respect to all variables for the most interesting specifications of Table 1. It is clear that the marginal effects are at least as reasonable in the most general model with autocorrelation as in other specifications. For instance, the linear model implies an unplausible but statistically significant complementarity between taxes and savings.

In general, we note that payments to the Social Pension Insurance System do not affect savings[4] — clearly because the average income of

[4] Taking out this variable in equation 7 did not change the log-likelihood significantly (-262.671).

Table 2

Savings function of higher-income white-collar German households (1978): Elasticities and [partial derivatives]

		1	2	4	7
		Linear	Log	Box-Cox	Box-Cox + Het. + Correlation
Payments to the Social Pens. Insur.	WHSPI	0.042 [0.046]	-0.011 [-0.011]	-0.018 [-0.018]	0.025 [0.025]
Household Size	WHPER	0.103 [20.500]	-0.166 [30.200]	-0.151 [-27.400]	-0.100 [-18.500]
Age of Household's head	WHAGE	-0.270 [-4.070]	-0.399 [-5.520]	-0.391 [-5.400]	-0.356 [-5.010]
Average return on household portfolio	WHINT	-0.649 [-9741.300]	-0.054 [-748.000]	-0.100 [-1370.000]	0.329 [4608.500]
Income and capital taxes	WHTAX	0.655 [0.448]	-0.359 [-0.226]	-0.309 [-0.194]	-0.221 [-0.141]
Income (before taxes)	WHGIN	1.080 [0.128]	2.640 [0.286]	2.560 [-0.278]	2.340 [0.258]

our group lies above the income contribution limit and consequently contributions vary very little. From the influence of household size we conclude that one additional person reduces savings by about DM 20 — a reasonable result because we know that this additional person does not earn. A reduction of savings with the aging of the household's head is compatible with the life cycle model. We also note the convincing result that income taxes reduce savings. The value of the marginal propensity to save (the complement of which is the marginal consumption rate) is credible and the respective value of the elasticity coefficient makes savings a superior good. A positive autocorrelation gives evidence of an imitation behavior ("keeping up with the Jones") which is very typical for this category of households.

CONCLUSION

We have pointed out that the structure of correlation among residuals could be exploited to test behavioral assumptions and noted that the formulation of required "residue impact criteria" generally implied a direct ordering of observations that only exceptionally made use of whatever degree of natural ordering the data might have possessed. We provided an example of directly ordered data where information in the residuals made behavioral sense and improved the regression results of a Box-Cox specification controlled for heteroskedasticity of a very general form.

BIBLIOGRAPHY

Blum, U.C., 1986, Raumwirkungen des Gesetzlichen Rentenversicherungsbudget, Karlsruher Beiträge zur Wirtschaftspolitik und Wirtschaftsforschung 13, v. Loeper Verlag, Karlsruhe.

Bolduc, D., 1985, Deux procédures d'estimation en présence d'autocorrélation spatiale dans les résidus, Publication #443, Centre de recherche sur les transports, Université de Montréal.

Gaudry, M.I.J. and M.G. Dagenais, 1979, Heteroskedasticity and the use of Box-Cox transformations, Economics Letters 2/3, 225-229.

King, M.L. and M.A. Evans, 1985, The Durbin-Watson test and cross-sectional data, Economics Letters 18, 31-34.

Liem, T.C., M.G. Dagenais and M.J.I. Gaudry, 1983, L-1.1, A program for Box-Cox transformations with heteroskedastic and autoregressive residuals, Publication #301, Centre de recherche sur les transports, Université de Montréal.

Statistisches Bundesamt, 1982, Einkommens- und Verbrauchsstichprobe 1978, No.4, Stuttgart-Mainz.